NOTES 04: CATEGORICAL VARIABLES

Stat 120 | Fall 2025

Prof Amanda Luby

1 Categorical Variables

Categorical variables are best summarized with a frequency table and visualized using a barplot. When we want to summarize a categorical variable with a single number, we often use a proportion.

> Proportion

When we have two categorical variables, we often use a two-way table to summarize them at the same time (also called the joint distribution). We might also care about the marginal distribution (the margins) or conditional distribution (a specific row/column).

Example: Below is the two-way table for our class representing the answers to "Have you taken a CS class before?" and whether the "Environmental Issues" interest box was checked.

	No	Yes - CS 111	Yes - Other
FALSE	10	9	3
TRUE	6	4	1

a. What is the marginal distribution of environmental interest?

b. What is the conditional distribution of environmental interest among those who have not taken a CS course?

c. What is the conditional distribution of prior CS courses among those who are not interested in environmental issues?

d. What is the proportion of students who have taken a prior CS course?

e. Does interest in environmental issues appear to be independent of prior CS experience?

2 Quantitative Variables

Quantitative variables are best visualized with a histogram or dotplot (depending on sample size)

When describing quantitative variables, we typically care most about the shape and center. When we want to summarize a quantitative variable with a single number, we often choose the mean, median, or mode.

Skewed Right	Symmetric	Skewed Left

.

There are various ways to describe the center of the distribution. The three most common are:

Mean

Median

Mode